Personality Expression in Spoken Dialogue Systems: From Text to Speech

Kenta Yamamoto and **Kazunori Komatani**
SANKEN, University of Osaka, Japan
{kentayamamoto,komatani}@sanken.osaka-u.ac.jp

Abstract

A consistent personality in a spoken dialogue system enhances the naturalness and friendliness of interactions. However, users may not accurately perceive all the personality traits that the system attempts to express. This study aims to identify which traits are most reliably perceived by users. We first analyzed third-party personality ratings of a dialogue corpus using principal component and factor analyses to uncover the underlying dimensions of user perception. We then conducted experiments under both text-only and speech-based dialogue conditions to evaluate how effectively each trait could be perceived. Crowd-sourced ratings showed that a trait concerning Extraversion and Openness can be reliably perceived through text alone, whereas accurate perception of the other traits requires speech-related features such as speech rate, backchannels, fillers, and turn-taking pause duration. These findings suggest that, rather than attempting to express all Big Five traits, focusing on a subset aligned with users' perceptual tendencies enables more effective and expressive personality design in spoken dialogue systems.

1 Introduction

Maintaining consistent behavior in a dialogue system is essential for achieving natural interaction and sustaining user trust. When system responses are inconsistent or contradictory, users can become confused or disengaged, reducing the perceived reliability of the system. Such behavioral consistency is often interpreted by users as the system's "personality," which helps them perceive the system as a coherent and trustworthy conversational partner (Nass et al., 1995).

A number of studies have explored ways to express personality in dialogue systems, for example by controlling the linguistic style of utterances (Mairesse and Walker, 2011; Saha et al., 2022; Shao et al., 2023; Caron and Srivastava, 2023) or by manipulating acoustic and prosodic features (Yamamoto et al., 2023b). Most of these studies adopt the Big Five model of personality—Extraversion (E), Openness (O), Agreeableness (A), Conscientiousness (C), and Neuroticism (N)—as it is one of the most widely used and empirically supported frameworks in psychology (Goldberg, 1990; Costa and McCrae, 1992). While the Big Five offers a comprehensive and generalizable description of human personality, it may be challenging for dialogue systems to express all five traits effectively within limited conversational contexts (Caron and Srivastava, 2023).

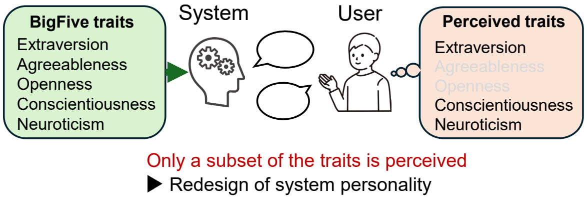

Figure 1: Perception of Big Five personality traits in first-time interactions

However, it remains difficult for users to accurately perceive all traits, even when the system displays a wide range of them (Figure 1). Some studies have suggested that only a limited subset of traits may be perceptible in such short interactions. In particular, third-party evaluations of dialogue participants have shown that Extraversion can be relatively well identified from short exchanges, whereas the remaining Big Five traits are much harder to infer (Komatani et al., 2023; Caron and Srivastava, 2023). For example, Openness, which reflects intellectual curiosity and interest in new experiences, is difficult to judge from a brief casual conversation because such tendencies are rarely expressed explicitly. Furthermore, the perceptibility of personality traits may vary across communication modalities: speech-based dialogue provides prosodic and paralinguistic cues

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 34–42
February 26–March 1, 2026. ©2026 Association for Computational Linguistics

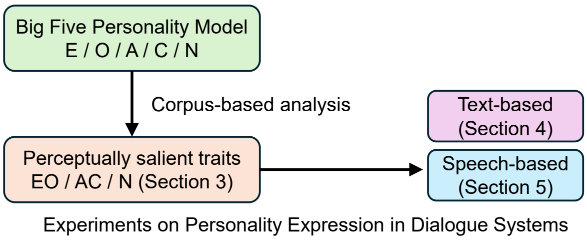
Big Five Personality Model
E / O / A / C / N
Corpus-based analysis
Perceptually salient traits
EO / AC / N (Section 3)
Text-based
(Section 4)
Speech-based
(Section 5)
Experiments on Personality Expression in Dialogue Systems

Figure 2: Overview of the research framework

that text-based chat lacks, potentially altering how personality is perceived by others.

Our study aims to identify which personality traits are perceptible to users during first-time interactions and to explore how dialogue systems can effectively express those traits. Our research is guided by three key questions:

- Which dimensions of the Big Five are more likely to be perceived by users in first-encounter dialogues, as suggested by our corpus analysis? (Section 3)
- How many dimensions of personality expression are perceptually sufficient in text-based dialogue—specifically, is a reduced set of three dimensions clearer than representing all five? (Section 4)
- How does the perceptual clarity of personality expression differ between text-based and speech-based dialogue? (Section 5)

To address these questions, we conducted a three-step investigation (Figure 2). First, we analyzed a public dialogue corpus annotated with Big Five traits to identify which dimensions are easily confused or less distinguishable by third-party evaluators (Section 3). Second, we conducted an experiment using text-based dialogue to examine whether the identified subset of traits can be perceived more clearly than the full set of Big Five traits (Section 4). Finally, we ran a speech-based dialogue experiment to investigate how speech features can be used to express those traits more clearly (Section 5).

2 Related Work

A wide range of studies have explored methods for incorporating personality traits into dialogue systems. Early approaches used rule-based generation frameworks such as PERSONAGE (Mairesse and Walker, 2011) or persona-conditioned sequence-to-sequence models (Li et al., 2016; Oraby et al., 2018), which embedded speaker information into the decoder to produce personalized utterances. Other research leveraged role-playing or character-based dialogue data to enable consistent stylistic expression (Higashinaka et al., 2018), and several persona-oriented datasets have been released to facilitate this goal (Zhang et al., 2018; Yamashita et al., 2023).

The advent of large language models (LLMs) has greatly expanded the possibilities for modeling and generating personality-consistent utterances. Recent work has attempted to endow LLMs with stable persona traits through fine-tuning (Shao et al., 2023) or prompting techniques that inject personality-representative descriptions or episodes (Caron and Srivastava, 2023). However, despite these advances, reliably controlling personality expression in LLM-generated dialogue remains difficult.

Among various personality frameworks, the Big Five model is the most widely adopted for dialogue research because of its psychological validity and interpretability. Nonetheless, previous studies have consistently reported asymmetric expressiveness across the five traits. Extraversion is typically easy to convey—often reflected in verbosity, enthusiasm, or engagement—whereas traits such as Conscientiousness or Neuroticism are much harder to express through text alone (Lotfi et al., 2023; Han et al., 2024; Caron and Srivastava, 2023). Prompt-based approaches using trait-related adjectives (Jiang et al., 2024) similarly found that only certain traits (especially Extraversion) are reliably recognized, underscoring fundamental limits in purely linguistic expression.

One key factor behind this limitation lies in the interaction context. In first-time or short dialogues, only sparse cues are available for inferring stable personality traits, which constrains both expression and perception. Moreover, as the Big Five was originally designed for self-assessment, applying it to perceived personality in brief interactions may introduce discrepancies between intended traits and user impressions (Komatani et al., 2023).

Beyond linguistic content, several studies have emphasized the importance of behavioral and prosodic factors—such as timing, intonation, and speech rate—in shaping personality impressions (Yamamoto et al., 2023b). These multimodal cues can enrich perceived expressiveness, particularly for traits like Neuroticism or Agreeableness that rely heavily on affective tone.

In light of these challenges, the present study takes a perception-oriented perspective. Rather than assuming that all Big Five traits are equally perceivable, we identify which dimensions are most salient in short, first-encounter dialogues and propose a simplified, empirically grounded three-trait configuration. We then demonstrate that these traits can be effectively expressed through both linguistic and prosodic cues, offering a practical direction for building personality-aware dialogue systems.

3 Empirical Selection of Personality Traits

Previous studies have shown that not all traits are equally perceptible in short or first-time conversations (Komatani et al., 2023). Expressing all five traits may therefore introduce unnecessary complexity and reduce the clarity of the system's personality.

In this section, we identify a simplified subset of personality traits that are reliably perceived by users. Based on third-party impressions from a multimodal Japanese dialogue corpus, rather than self-assessments, we analyze how each trait is recognized. This approach reflects the practical goal of dialogue systems: what matters is not self-expression but how the system is perceived.

To reveal perceptually salient traits, we apply principal component analysis (PCA) and exploratory factor analysis (EFA) to the third-party ratings. These analyses reduce dimensionality and uncover psychologically interpretable clusters of traits, forming the basis for a more controllable and perceptible personality model.

3.1 Dataset

We used the multimodal Japanese dialogue corpus *Hazumi* (Komatani and Okada, 2021) for our analysis. This corpus consists of Wizard-of-Oz (WoZ) style conversations between a human operator controlling an agent ("Mei") and human participants engaging in casual topics. A total of 155 dialogues were selected from four sub-corpora: 1911[1], 2010[2], 2012[3], and 2105[4]. These include both in-person and online interactions. Although the system utterances were manually operated, the dialogues maintained natural conversational flow and variability, making them suitable for analyzing perceived personality.

Table 1: Principal component analysis (PCA) on third-party ratings of the Big Five traits.

PC	1	2	3	4	5
Variance	0.49	0.22	0.15	0.10	0.04
Cumulative	0.49	0.71	**0.86**	0.96	1.00

Personality ratings were provided by third-party evaluators who watched the recorded videos. Each evaluator rated the perceived personality of participants using the Japanese Ten Item Personality Inventory (TIPI-J) (Oshio et al., 2012), a validated 10-item questionnaire measuring the Big Five traits on a 7-point Likert scale. Five annotators independently rated each dialogue, and their scores were averaged to mitigate individual bias. This third-person evaluation approach captures impressions similar to how users perceive a conversational partner in first-time encounters.

3.2 Analysis

To identify which personality dimensions are perceptually salient, we conducted a two-stage analysis combining principal component analysis (PCA) and exploratory factor analysis (EFA). We applied PCA followed by Varimax-rotated EFA to identify latent perceptual dimensions.

Principal Component Analysis (PCA) PCA was first applied to examine the dimensionality of the third-party ratings and to estimate how much of the total variance in perceived personality could be explained by fewer latent components. This step provided an empirical basis for dimensionality reduction: if a small number of components explained most of the variance, it would indicate that not all five Big Five traits are clearly distinguished by observers.

As shown in Table 1, the first three components together accounted for 86% of the total variance. This indicates that three orthogonal dimensions are sufficient to capture most of the variation in perceived personality, while additional components contribute only marginally. Accordingly, we limited the subsequent factor analysis to three factors.

Exploratory Factor Analysis (EFA) EFA with Varimax rotation was then applied to interpret the structure of the three latent factors and to clarify how specific traits cluster perceptually. Varimax rotation was chosen to enhance interpretability by reducing cross-loadings among traits.

As shown in Table 2, three perceptually coherent

[1] https://github.com/ouktlab/Hazumi1911/
[2] https://github.com/ouktlab/Hazumi2010/
[3] https://github.com/ouktlab/Hazumi2012/
[4] https://github.com/ouktlab/Hazumi2105/

Table 2: Exploratory factor analysis (EFA) on third-party ratings of the Big Five traits.

	Factor		
	1	2	3
E	**0.88**	0.14	0.02
O	**0.76**	0.10	0.50
A	−0.01	**0.56**	0.19
C	0.12	**0.43**	**0.73**
N	−0.34	−0.13	**−0.51**

Table 3: Spearman's rank correlation coefficients among the Big Five traits (third-party ratings).

		Annotation rating of Big Five traits				
		E	O	A	C	N
Rating	E	–	**0.68**	0.19	0.05	−0.32
	O	–	–	0.22	0.38	−0.51
	A	–	–	–	**0.34**	−0.20
	C	–	–	–	–	−0.39

clusters emerged. Factor 1 primarily corresponds to *Extraversion* and *Openness*, both related to expressiveness and sociability. Factor 2 represents a blend of *Agreeableness* and *Conscientiousness*, reflecting cooperative and responsible impressions. Factor 3 corresponds mainly to *Neuroticism*, associated with emotional instability. These findings suggest that in short, first-encounter dialogues, personality impressions can be effectively represented by three interpretable dimensions rather than all five theoretical traits.

3.3 Results and Interpretation

The factor analysis revealed three perceptually distinct dimensions that summarize how observers infer personality in short dialogues. To further confirm these latent structures, we examined pairwise correlations among the original Big Five ratings (Table 3).

EO: Extraversion and Openness showed both high factor loadings and the strongest positive correlation ($r = 0.68$). This indicates that enthusiasm, curiosity, and verbal fluency are perceived jointly as a single expressive dimension—qualities that are easily observable even in short interactions.

AC: Agreeableness and Conscientiousness moderately co-varied ($r = 0.34$) and shared common factor loadings, representing politeness, cooperativeness, and responsibility. These traits jointly shape impressions of warmth and dependability in conversation partners.

N: Neuroticism was negatively correlated with Conscientiousness ($r = -0.39$) and Openness ($r = -0.51$), consistent with its negative loading in the factor analysis. This suggests that emotional instability is perceived in opposition to organized or composed behavior, and that stability cues—such as calmness and coherence—underlie this dimension.

Taken together, the PCA, factor analysis, and correlation patterns consistently indicate that the Big Five traits are perceptually compressed into three composite dimensions—EO (Extraversion + Openness), AC (Agreeableness + Conscientiousness), and N (Neuroticism). This perceptual compression suggests that in first-time interactions, users interpret system personality primarily through a limited set of salient social cues rather than all five theoretical dimensions. Accordingly, we hypothesize that dialogue systems expressing these three empirically grounded traits will achieve clearer and more consistent personality impressions than those attempting to represent all five Big Five traits. The following experiments (Sections 4–5) test this hypothesis in both text- and speech-based dialogue settings.

4 Personality Expression in Text-Based Dialogue Systems

This section examines how effectively the three perceptually salient personality dimensions—EO, AC, and N—can be expressed in dialogue systems compared with representing all five Big Five traits. We focus first on text-based interactions, where personality must be conveyed solely through linguistic cues, and later extend the analysis to speech-based dialogues that include prosodic information.

4.1 Experimental Design and Procedure

An overview of the overall procedure is shown in Figure 3. The experimental dialogues were derived from the Hazumi1911 corpus. Ten dialogues were selected, and ten consecutive turns on a single topic were extracted from each transcript to ensure topical coherence.

For each dialogue, we generated both text- and speech-based versions by systematically controlling three composite personality traits—EO (Extraversion + Openness), AC (Agreeableness + Conscientiousness), and N (Neuroticism)—each at two levels ("high" and "low"). This resulted in eight LLM-generated personality conditions plus the original corpus version, yielding nine conditions per dialogue and a total of 90 samples.

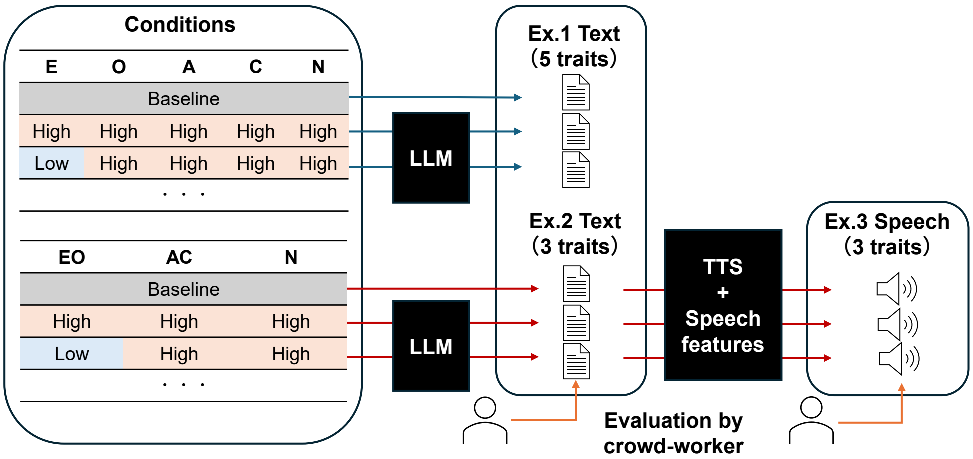

Figure 3: Overview of the experiments.

Table 4: Prompt used for LLM-based utterance generation.

You will engage in small talk with the user as a female
character named Mei. You have {PERSONALITY_WORDS}
personality. Please ensure your utterances reflect this
personality. Keep your responses concise and follow
the flow of the conversation while maintaining consistency
in your personality throughout the dialogue. Now, begin.
{Dialogue_History}

System utterances were replaced with LLM-generated ones while user utterances remained unchanged. This process was repeated until all system turns were substituted, resulting in fully controlled dialogue samples under each personality condition.

Evaluators were recruited via a crowdsourcing platform[5]. All participants were native Japanese speakers.

Evaluators rated the system's personality using the Japanese Ten Item Personality Inventory (TIPI-J) (Oshio et al., 2012), a validated ten-item scale measuring the Big Five traits on a seven-point Likert scale. Each Big Five trait is represented by two items, and the mean of the two formed the trait score. These items are formulated as statements describing personality concepts, thereby facilitating evaluators' understanding of the target concepts. We then computed EO as the mean of Extraversion and Openness, AC as the mean of Agreeableness and Conscientiousness, and N directly from the Neuroticism score.

4.2 Utterance Generation Using LLM

We employed GPT-4 (version gpt-4-0613) as the large language model (LLM), with the temperature fixed at 0 to ensure deterministic output. Personality traits were controlled through prompt-based specification without fine-tuning or example-based conditioning, ensuring zero-shot generation.

The system character was named "Mei", consistent with the agent used in the corpus. For each dialogue, all system utterances in the original transcripts were replaced with LLM-generated responses, while user utterances were kept unchanged to preserve conversational naturalness. The model generated each system turn based on the dialogue history consisting of alternating user–system exchanges, using the most recent user utterance as input.

Table 4 shows the prompt template used for utterance generation. Personality control was achieved by substituting the placeholder {PERSONALITY_WORDS} with adjectives corresponding to each target trait and intensity level. These adjectives were selected with reference to prior studies on lexical representations of the Big Five personality traits (Goldberg, 1990).

4.3 Experiment 1: Expression of Individual Big Five Traits in Text-Based Dialogue

Experiment 1 investigated whether each of the five Big Five personality traits could be individually expressed and perceived through text-based dialogue generated by the LLM. In this experiment, 100 workers each rated ten dialogues. This experiment serves as a baseline for evaluating the perceptual distinctiveness of the proposed three-dimensional model (EO, AC, N) introduced in later sections.

As shown in Table 5, all five traits exhibited clear differentiation between the high and low control conditions, indicating that the LLM successfully

[5] https://crowdworks.jp/

Table 5: Experiment 1 (Text-Based Dialogue): Big Five rating scores (7-point scale)

		Mean rating scores (Std)				
Traits	Condition	E	O	A	C	N
E	High	5.0 (1.4)*	4.2 (1.3)*	5.4 (1.2)*	4.7 (1.0)	2.9 (1.0)*
	Low	3.2 (1.4)*	3.1 (1.3)*	3.3(1.9)*	4.4 (1.1)	3.9 (1.5)*
O	High	4.7 (1.0)*	4.3 (1.1)*	5.6 (1.2)*	5.5 (0.9)*	2.3 (0.8)*
	Low	3.4 (1.1)*	3.1 (1.2)*	4.0 (1.7)*	4.4 (1.1)*	3.4 (1.2)*
A	High	4.9 (1.1)*	4.5 (1.1)*	6.1 (0.7)*	5.4 (0.8)*	2.2 (0.7)*
	Low	3.9 (1.4)*	3.1 (1.1)*	2.6 (1.9)*	4.0 (1.3)*	3.9 (1.3)*
C	High	4.5 (1.1)*	3.8 (1.0)*	5.3 (1.4)*	5.1 (1.0)*	2.5 (0.9)*
	Low	3.1 (1.1)*	2.7 (1.2)*	3.2 (1.3)*	3.1 (1.2)*	4.4 (1.3)*
N	High	4.6 (1.2)*	4.1 (1.1)	4.8 (1.9)*	4.9 (1.1)	3.1 (1.4)*
	Low	4.1 (1.3)*	3.8 (1.1)	5.5 (1.3)*	5.0 (1.1)	2.9 (1.1)*

Welch's two-sided t-test (*$p < 0.05$)

Table 6: Experiment 1 (Text-Based Dialogue): Spearman's rank correlation coefficients between the specified personality control settings and Big Five rating scores.

		Rated Traits				
		E	O	A	C	N
Control	E	**0.27**	0.18	**0.24**	0.06	−0.13
	O	**0.22**	**0.23**	**0.21**	**0.21**	**−0.20**
	A	0.16	**0.24**	**0.39**	**0.27**	**−0.30**
	C	**0.20**	0.19	**0.25**	**0.31**	**−0.29**
	N	0.08	−0.05	−0.08	−0.03	0.09

Table 7: Experiment 2 (Text-Based Dialogue): personality rating scores (7-point scale)

		Mean rating scores (Std)		
Traits	Condition	EO	AC	N
Corpus-based		3.8 (0.9)	4.5 (0.9)	3.6 (1.0)
EO	High	4.8 (1.0)*	4.9 (0.9)*	3.0 (1.0)
	Low	3.7 (0.9)*	3.8 (1.4) *	3.4 (1.1)
AC	High	4.3 (1.1)	5.0 (0.9)*	3.0 (1.1)
	Low	4.4 (1.1)	4.0 (1.4)*	3.4 (1.0)
N	High	4.4 (1.3)	4.4 (1.1)	3.4 (1.1)
	Low	4.3 (1.1)	4.6 (1.1)	3.1 (1.0)

Welch's two-sided t-test (*$p < 0.05$)

Table 8: Experiment 2 (Text-Based Dialogue): Spearman's rank correlation coefficients between the specified personality control settings and the rating scores.

		Rated Traits		
		EO	AC	N
Control	EO	**0.47**	**0.43**	−0.18
	AC	0.04	**0.36**	−0.13
	N	0.07	−0.08	0.14

generated text reflecting the intended personality manipulations. Extraversion and Openness showed the most pronounced changes, with higher values associated with increased perceived Agreeableness and Conscientiousness and decreased Neuroticism. This trend suggests that linguistic expressions conveying enthusiasm, curiosity, and engagement (e.g., active phrasing, inclusive language) simultaneously enhance impressions of warmth and reliability. In contrast, manipulations of Neuroticism produced smaller and less consistent differences, implying that emotional instability is less effectively conveyed through textual cues alone.

Table 6 lists the Spearman's rank correlation coefficients between the specified control trait and each rated trait. Positive correlations along the diagonal indicate that the target trait was generally perceived as intended. However, off-diagonal correlations reveal notable cross-trait effects—for instance, manipulations of Extraversion or Agreeableness also increased ratings of Openness and Conscientiousness. This suggests that linguistic cues such as friendliness, enthusiasm, or engagement simultaneously influence multiple trait impressions, making it difficult to isolate all five traits purely through text.

4.4 Experiment 2: Expression of Three Composite Traits in Text-Based Dialogue

Experiment 2 evaluated whether the three composite personality traits derived from the corpus analysis—EO (Extraversion + Openness), AC (Agreeableness + Conscientiousness), and N (Neuroticism)—can be effectively expressed and perceived through text-based dialogue. This experiment directly tests the three-dimensional model proposed in Section 3 using linguistic cues alone. In this experiment, 90 workers each rated ten dialogues.

Table 7 shows the mean and standard deviation of the personality ratings for each condition. A Welch's two-sided t-test revealed significant differences ($p < .05$) between the high and low conditions for each target trait, confirming that the intended manipulations were successfully reflected in perceived personality. However, some cross-trait effects were observed—for instance, increasing EO slightly elevated AC ratings—indicating partial overlap in linguistic cues such as friendliness and engagement.

Table 8 presents the Spearman's rank correlation coefficients between control settings and corresponding evaluation scores. Control levels were coded as high (1), corpus-based (0), and low (–1). The results show strong positive correlations between EO control and both EO and AC ratings, whereas correlations for Neuroticism were weaker. These findings suggest that EO and AC are the most salient dimensions in text-based dialogue, but their perceptual separation is limited because both rely on similar linguistic indicators of sociability and cooperativeness. In contrast, N remains difficult to convey due to the absence of paralinguistic signals.

4.5 Discussion

The three personality traits (EO, AC, N) provided clearer and more consistent personality expression than the full Big Five. EO and AC were clearly reflected in evaluator ratings, suggesting that expressiveness and interpersonal reliability can be conveyed through lexical choice, phrasing, and tone.

Some overlap was observed between EO and AC, indicating that linguistic cues for sociability and cooperativeness partially overlap—e.g., friendly or engaging phrasing. Nevertheless, these traits remained distinguishable, demonstrating that the three-trait configuration enables reasonably independent control.

In contrast, Neuroticism (N) showed weaker correspondence between control conditions and perceived ratings, implying that emotional instability is difficult to convey without nonverbal cues such as hesitation or prosody. This finding supports the view that N requires multimodal reinforcement.

Comparing the five-trait and three-trait models highlights a key advantage of simplification. Directly controlling individual Big Five traits produced strong cross-trait correlations, meaning that altering one trait often influenced perceptions of others. In contrast, the three-trait model reduced this interference: EO and AC could be expressed more independently, and evaluator judgments were more consistent. Thus, reducing representational dimensionality to empirically salient traits improves both controllability and perceptual clarity.

Overall, these results indicate that traits associated with linguistic engagement (EO and AC) can be effectively manipulated through text alone, whereas traits linked to emotion (N) benefit from additional modalities.

Table 9: Speech feature settings in speech-based dialogue conditions

	Condition	
Traits	High	Low
EO	Backchannels inserted	None
AC	Long pause (3.0 s)	Short pause (0.5 s)
N	Fillers added; variable speech rate	None

5 Experiment 3: Expression of Three Composite Traits in Speech-Based Dialogue

This experiment investigated whether personality traits can be more effectively expressed through spoken dialogue than through text alone. Building upon the same dialogue content used in the text-based condition, we synthesized complete spoken dialogues by combining system-generated utterances with user utterances extracted from the corpus. Speech synthesis was performed using the female voice provided by the VoiceText text-to-speech (TTS) engine.

5.1 Settings

For each dialogue, all ten system utterances were synthesized according to the same linguistic content used in Experiment 1. Each personality trait—EO, AC, and N—was assigned a "high" or "low" condition, resulting in six controlled variations in total. Table 9 summarizes the specific speech features applied to each trait.

The control of behaviors in this experiment was based on the previous study (Yamamoto et al., 2023a). For the high EO condition, backchannels such as "hai" were inserted at 3- and 5-second intervals during user speech, while none were added for the low EO condition. For high AC, the system's turn-taking pause was extended from the corpus-based value of 1.0 s to 3.0 s; for low AC, it was shortened to 0.5 s. For high N, fillers such as "e–" and "sono–" were inserted at the beginning of each system utterance, and the speech rate alternated between 80% (slower) and 110% (faster) of the corpus default. No fillers or rate changes were applied for low N.

In this experiment, 180 workers each rated five dialogues using the same evaluation criteria as in the experiment described in Section 4.4. Dialogue samples were presented to evaluators as approximately 1.5-minute videos. To avoid visual bias, the videos consisted of a black background with only two icons representing the user and the system.

Table 10: Experiment 3 (Speech-Based Dialogue): personality rating scores (7-point scale)

		Mean rating scores (Std)		
Traits	Condition	EO	AC	N
Corpus-based		4.2 (0.9)	4.9 (1.0)	2.7 (0.9)
EO	High	4.6 (0.9)*	4.5 (1.2)	3.2 (1.1)
	Low	4.0 (0.9)*	4.2 (1.2)	3.3 (1.2)
AC	High	4.4 (0.9)	4.8 (1.0)*	3.1 (1.1)
	Low	4.3 (1.0)	4.0 (1.3)*	3.4 (1.1)
N	High	4.1 (0.9)	4.1 (1.2)	3.6 (1.2)*
	Low	4.5 (0.9)	4.5 (1.2)	3.0 (1.0)*

Welch's two-sided t-test (*$p < .05$)

5.2 Results

Table 10 presents the mean and standard deviation of personality rating scores under each condition. A Welch's two-sided t-test revealed significant differences ($p < .05$) only when the evaluated trait matched the controlled trait, indicating successful manipulation of the intended dimensions. In contrast to the text-based condition, cross-trait influences (e.g., EO affecting AC) were minimal.

Table 11 shows the Spearman's rank correlation coefficients between control settings and corresponding evaluation scores. Control conditions were encoded as high (1), corpus-based (0), and low (–1). The results indicate positive correlations between each controlled trait and its respective ratings, with notably higher values for Neuroticism compared to the text-based condition.

5.3 Discussion

The results demonstrate that incorporating prosodic and timing cues in speech allows for clearer and more independent control of personality expression. In particular, the EO and AC traits were perceived consistently across evaluators, while the expression of N—difficult to convey through text—became significantly more salient through speech cues such as fillers and speech-rate variation.

Compared to the text-only condition, the correlations between controlled and perceived traits were stronger for N and more distinct overall, indicating reduced cross-trait interference. These findings suggest that multimodal cues, including backchannels, timing, and speech rate, play an essential role in the perceptual realization of personality, complementing linguistic information. Thus, spoken dialogue enables a richer and more distinguishable expression of personality traits than text-based dialogue alone.

On the other hand, using speech also introduces influences on personality impressions through TTS. Therefore, validation using different TTS systems will likely be necessary in future works.

Table 11: Experiment 3 (Speech-Based Dialogue): Spearman's rank correlation coefficients between control settings and personality ratings

		Correlation coefficients		
		EO	AC	N
Control	EO	**0.33**	0.09	−0.05
	AC	0.04	**0.33**	−0.15
	N	−0.19	−0.16	**0.28**

6 Conclusion

This study examined how personality traits can be expressed and perceived in dialogue systems, focusing on first-time interactions where impressions form rapidly.

Analysis of third-party ratings in a multimodal Japanese dialogue corpus showed that three composite dimensions—EO, AC, and N—explain most variance in perceived personality. This three-trait model, rather than replacing the Big Five, provides an empirically grounded simplification reflecting how users perceive system personalities in brief exchanges.

Text-based experiments with large language models confirmed that the three-trait representation yielded clearer and more consistent impressions than directly manipulating all five traits. EO and AC were reliably expressed through linguistic features such as politeness and dialogue flow, while N was harder to convey without prosodic cues. Simplifying personality to three salient dimensions improved both consistency and interpretability.

Speech-based experiments further showed that prosodic and paralinguistic cues enhanced expressiveness, especially for N, which became perceptually distinct with fillers, backchannels, and variable speech rate. These results underscore the role of speech in conveying emotional and interpersonal nuance.

In sum, modeling perceptually salient traits and integrating linguistic and prosodic cues offers a practical, cognitively plausible framework for personality expression in dialogue systems. Future work should explore adaptive multimodal control incorporating visual and contextual signals, and examine cross-linguistic and longitudinal aspects of personality perception. Furthermore, as our study was conducted in Japanese, the results should be validated in other languages and cultural contexts.

Acknowledgement

This work was supported by Grant-in-Aid for Research Activity Start-up Grant Number 23K20005 and Grant-in-Aid for Early-Career Scientists Grant Number 24K20839.

References

Graham Caron and Shashank Srivastava. 2023. Manipulating the perceived personality traits of language models. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 2370–2386.

Paul T. Costa and Robert R. McCrae. 1992. Normal personality assessment in clinical practice: The NEO personality inventory. *Psychological Assessment*, 4(1):5–13.

Lewis R. Goldberg. 1990. An alternative "description of personality": the big-five factor structure. *Personality and Social Psychology*, 59(6):1216–1229.

Ji-Eun Han, Jun-Seok Koh, Hyeon-Tae Seo, Du-Seong Chang, and Kyung-Ah Sohn. 2024. Psydial: Personality-based synthetic dialogue generation using large language models. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 13321—-13331.

Ryuichiro Higashinaka, Masahiro Mizukami, Hidetoshi Kawabata, Emi Yamaguchi, Noritake Adachi, and Junji Tomita. 2018. Role play-based question-answering by real users for building chatbots with consistent personalities. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 264–272.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics (NAACL)*, pages 3605–3627.

Kazunori Komatani and Shogo Okada. 2021. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Kazunori Komatani, Ryu Takeda, and Shogo Okada. 2023. Analyzing differences in subjective annotations by participants and third-party annotators in multimodal dialogue corpus. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 1688–1692.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 994–1003.

Ehsan Lotfi, Maxime De Bruyn, Jeska Buhmann, and Walter Daelemans. 2023. PersonalityChat: Conversation distillation for personalized dialog modeling with facts and traits. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 353–371.

Francois Mairesse and Marilyn A. Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3):455–488.

Clifford Nass, Youngme Moon, B.J.Fogg, Byron Reeves, and D.Christopher Dryer. 1995. Can computer personalities be human personalities? *Human-Computer studies*, 43:223–239.

Shereen Oraby, Lena Reed, Shubhangi Tandon, Sharath T.S., Stephanie Lukin, and Marilyn Walker. 2018. Controlling personality-based stylistic variation with neural natural language generators. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 180–190.

Atsushi Oshio, Shingo Abe, and Pino Cutrone. 2012. Development, reliability, and validity of the japanese version of ten item personality inventory (TIPI-J). *The Japanese Journal of Personality*, 21(1):42–52.

Sougata Saha, Souvik Das, and Rohini Srihari. 2022. Stylistic response generation by controlling personality traits and intent. In *Proceedings of the Workshop on NLP for Conversational AI*, pages 197–211.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 13153–13187.

Kenta Yamamoto, Koji Inoue, and Tatsuya Kawahara. 2023a. Character expression for spoken dialogue systems with semi-supervised learning using variational auto-encoder. *Computer Speech Language*, 79:101469–101469.

Kenta Yamamoto, Koji Inoue, and Tatsuya Kawahara. 2023b. Character expression of a conversational robot for adapting to user personality. *Advanced Robotics*, 38(4):256–266.

Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 852–861.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2204–2213.